Chapter 11  
Specification Error Analysis

The specification of a linear regression model consists of a formulation of the regression relationships and of statements or assumptions concerning the explanatory variables and disturbances. If any of these is violated, e.g., incorrect functional form, incorrect introduction of disturbance term in the model etc., then specification error occurs. In narrower sense, the specification error refers to explanatory variables.

The complete regression analysis depends on the explanatory variables present in the model. It is understood in the regression analysis that only correct and important explanatory variables appears in the model. In practice, after ensuring the correct functional form of the model, the analyst usually has a pool of explanatory variables which possibly influence the process or experiment. Generally, all such candidate variables are not used in the regression modeling but a subset of explanatory variables is chosen from this pool.

While choosing a subset of explanatory variables, there are two possible options:

1. In order to make the model as realistic as possible, the analyst may include as many as possible explanatory variables.
2. In order to make the model as simple as possible, one may include only fewer number of explanatory variables.

In such selections, there can be two types of incorrect model specifications.

1. Omission/exclusion of relevant variables.
2. Inclusion of irrelevant variables.

Now we discuss the statistical consequences arising from the both situations.

1. Exclusion of relevant variables:

In order to keep the model simple, the analyst may delete some of the explanatory variables which may be of importance from the point of view of theoretical considerations. There can be several reasons behind such decisions, e.g., it may be hard to quantify the variables like taste, intelligence etc. Sometimes it may be difficult to take correct observations on the variables like income etc.
Let there be $k$ candidate explanatory variables out of which suppose $r$ variables are included and $(k-r)$ variables are to be deleted from the model. So partition the $X$ and $\beta$ as

$$X_{nk} = \begin{pmatrix} X_1 & X_2 \\ \end{pmatrix}_{n \times (k-r)} \text{ and } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \end{pmatrix}_{(r \times 1) \times (k-r \times 1)}.$$

The model $y = X\beta + \varepsilon$, $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2 I$ can be expressed as

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

which is called as **full model** or **true model**.

After dropping the $r$ explanatory variable in the model, the new model is

$$y = X_1\beta_1 + \delta$$

which is called as **misspecified model** or **false model**.

Applying OLS to the false model, the OLSE of $\beta_1$ is

$$b_{1F} = (X'_1X_1)^{-1}X'_1y.$$ 

The estimation error is obtained as follows:

$$b_{1F} = (X'_1X_1)^{-1}X'_1(X_1\beta_1 + X_2\beta_2 + \varepsilon)$$

$$= \beta_1 + (X'_1X_1)^{-1}X'_1X_2\beta_2 + (X'_1X_1)^{-1}X'_1\varepsilon$$

$$b_{1F} - \beta_1 = \theta + (X'_1X_1)^{-1}X'_1\varepsilon$$

where $\theta = (X'_1X_1)^{-1}X'_1X_2\beta_2$.

Thus

$$E(b_{1F} - \beta_1) = \theta + (X'_1X_1)^{-1}E(\varepsilon)$$

$$= \theta$$

which is a linear function of $\beta_2$, i.e., the coefficients of excluded variables. So $b_{1F}$ is biased, in general.

The bias vanishes if $X'_1X_2 = 0$, i.e., $X_1$ and $X_2$ are orthogonal or uncorrelated.

The mean squared error matrix of $b_{1F}$ is

$$MSE(b_{1F}) = E((b_{1F} - \beta_1)(b_{1F} - \beta_1)^{'})$$

$$= E\left[ \theta\theta' + \theta\varepsilon'X_1(X'_1X_1)^{-1} + (X'_1X_1)^{-1}X'_1\varepsilon\theta' + (X'_1X_1)^{-1}X'_1\varepsilon\varepsilon'X_1(X'_1X_1)^{-1} \right]$$

$$= \theta\theta' + 0 + 0 + \sigma^2(X'_1X_1)^{-1}X'_1IX_1(X'_1X_1)^{-1}$$

$$= \theta\theta' + \sigma^2(X'_1X_1)^{-1}.$$
So efficiency generally declines. Note that the second term is the conventional form of MSE.

The residual sum of squares is

\[ \hat{\sigma}^2 = \frac{SS_{res}}{n-r} = \frac{e'e}{n-r} \]

where \( e = y - X_1b_F = \tilde{H}_1y, \)

\[ \tilde{H}_1 = I - X_1(X_1'X_1)^{-1}X_1. \]

Thus

\[ \tilde{H}_1y = \tilde{H}_1(X_1\beta_1 + X_2\beta_2 + \varepsilon) = 0 + \tilde{H}_1(X_2\beta_2 + \varepsilon) = \tilde{H}_1(X_2\beta_2 + \varepsilon). \]

\[ y'\tilde{H}_1y = (X_1\beta_1 + X_2\beta_2 + \varepsilon)\tilde{H}_1(X_2\beta_2 + \varepsilon) = (\beta_1X_1'\tilde{H}_1X_2\beta_2 + \beta_1X_1'\tilde{H}_1\varepsilon + \beta_2X_2'\tilde{H}_1X_2\beta_2 + \beta_2X_2'\tilde{H}_1\varepsilon + \varepsilon'\tilde{H}_1X_2\beta_2 + \varepsilon'\tilde{H}_1\varepsilon). \]

\[ E(s^2) = \frac{1}{n-\nu} \left[ E(\beta_2X_2'\tilde{H}_1X_2\beta_2) + 0 + E(\varepsilon'\tilde{H}_1\varepsilon) \right] \]

\[ = \frac{1}{n-\nu} \left[ \beta_2X_2'\tilde{H}_1X_2\beta_2 + (n-\nu)\sigma^2 \right] \]

\[ = \sigma^2 + \frac{1}{n-\nu} \beta_2X_2'\tilde{H}_1X_2\beta_2. \]

Thus \( s^2 \) is a biased estimator of \( \sigma^2 \) and \( s^2 \) provides an over estimate of \( \sigma^2 \). Note that even if \( X_1'X_2 = 0 \), then also \( s^2 \) gives an overestimate of \( \sigma^2 \). So the statistical inferences based on this will be faulty. The \( t \)-test and confidence region will be invalid in this case.

If the response is to be predicted at \( x' = (x_1', x_2') \), then using the full model, the predicted value is

\[ \hat{y} = x'b = x'(X'X)^{-1}X'y \]

with

\[ E(\hat{y}) = x'\beta \]

\[ Var(\hat{y}) = \sigma^2 \left[ 1 + x'(X'X)^{-1}x \right]. \]

When subset model is used then the predictor is

\[ \hat{y}_1 = x'b_{1_F} \]

and then

---

*Econometrics | Chapter 11 | Specification Error Analysis | Shalabh, IIT Kanpur*
\[ E(\hat{y}_1) = \hat{x}_i (X_i' X_i)^{-1} X_i E(y) \]
\[ = \hat{x}_i (X_i' X_i)^{-1} X_i E(X_i \beta_1 + X_2 \beta_2 + \varepsilon) \]
\[ = \hat{x}_i (X_i' X_i)^{-1} X_i (X_i \beta_1 + X_2 \beta_2) \]
\[ = \hat{x}_i \beta_1 + \hat{x}_i (X_i' X_i)^{-1} X_i X_2 \beta_2 \]
\[ = \hat{x}_i \beta_1 + \hat{x}_i \theta. \]

Thus \( \hat{y}_i \) is a biased predictor of \( y \). It is unbiased when \( X_i' X_2 = 0 \). The MSE of predictor is
\[ MSE(\hat{y}_1) = \sigma^2 \left[ 1 + \hat{x}_i (X_i' X_i)^{-1} x_i \right] + \left( \hat{x}_i \theta - \hat{x}_i \beta_2 \right)^2. \]

Also
\[ Var(\hat{y}) \leq MSE(\hat{y}_1) \]

provided \( V(\hat{\beta}_2) - \beta_2 \beta_2^T \) is positive semidefinite.

2. Inclusion of irrelevant variables

Sometimes due to enthusiasm and to make the model more realistic, the analyst may include some explanatory variables that are not very relevant to the model. Such variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom \((n-k)\) and consequently the validity of inference drawn may be questionable. For example, the value of coefficient of determination will increase indicating that the model is getting better which may not really be true.

Let the true model be
\[ y = X \beta + \varepsilon, \ E(\varepsilon) = 0, V(\varepsilon) = \sigma^2 I \]
which comprise \( k \) explanatory variable. Suppose now \( r \) additional explanatory variables are added to the model and resulting model becomes
\[ y = X \beta + Z \gamma + \delta \]
where \( Z \) is a \( n \times r \) matrix of \( n \) observations on each of the \( r \) explanatory variables and \( \gamma \) is \( r \times 1 \) vector of regression coefficient associated with \( Z \) and \( \delta \) is disturbance term. This model is termed as false model.

Applying OLS to false model, we get
\[
\begin{pmatrix}
    b_f \\
    c_f
\end{pmatrix}
= \begin{pmatrix}
    X'X & X'Z \\
    Z'X & Z'Z
\end{pmatrix}^{-1}
\begin{pmatrix}
    X'y \\
    Z'y
\end{pmatrix}
\]

\[
\begin{pmatrix}
    X'X & X'Z \\
    Z'X & Z'Z
\end{pmatrix}
\begin{pmatrix}
    b_f \\
    c_f
\end{pmatrix}
= \begin{pmatrix}
    X'y \\
    Z'y
\end{pmatrix}
\]

\[
\Rightarrow X'Xb_f + X'Zc_f = X'y \quad (1)
\]
\[
Z'Xb_f + Z'Zc_f = Z'y \quad (2)
\]

where \( b_f \) and \( c_f \) are the OLSEs of \( \beta \) and \( \gamma \) respectively.

Premultiply equation (2) by \( X'Z(Z'Z)^{-1} \), we get
\[
X'Z(Z'Z)^{-1}Z'Xb_f + X'Z(Z'Z)^{-1}Z'Zc_f = X'Z(Z'Z)^{-1}Z'y. \quad (3)
\]

Subtracting equation (1) from (3), we get
\[
X'X - X'Z(Z'Z)^{-1}Z'X
\begin{pmatrix}
    b_f \\
    c_f
\end{pmatrix}
= X'y - X'Z(Z'Z)^{-1}Z'y
\]
\[
X'\left[ I - Z(Z'Z)^{-1}Z' \right]Xb_f = X'\left[ I - Z(Z'Z)^{-1}Z' \right]y
\]
\[
\Rightarrow b_f = (X'\widetilde{H}_z X)^{-1}X'\widetilde{H}_z y
\]

where \( \widetilde{H}_z = I - Z(Z'Z)^{-1}Z' \).

The estimation error of \( b_f \) is
\[
b_f - \beta = (X'\widetilde{H}_z X)^{-1}X'y - \beta
\]
\[
= (X'\widetilde{H}_z X)^{-1}X'\widetilde{H}_z (X\beta + \varepsilon) - \beta
\]
\[
= (X'\widetilde{H}_z X)^{-1}X'\widetilde{H}_z \varepsilon.
\]

Thus
\[
E(b_f - \beta) = (X'\widetilde{H}_z X)^{-1}X'\widetilde{H}_z E(\varepsilon) = 0
\]

so \( b_f \) is unbiased even when some irrelevant variables are added to the model.

The covariance matrix is
\[
V(b_f) = E(b_f - \beta)(b_f - \beta)^\top
\]
\[
= E\left[ (X'\widetilde{H}_z X)^{-1}X'\widetilde{H}_z \varepsilon'\widetilde{H}_z X(X'\widetilde{H}_z X)^{-1} \right]
\]
\[
= \sigma^2 (X'\widetilde{H}_z X)^{-1}X'\widetilde{H}_z \widetilde{H}_z X(X'\widetilde{H}_z X)^{-1}
\]
\[
= \sigma^2 (X'\widetilde{H}_z X)^{-1}.
\]
If OLS is applied to true model, then

\[ b_t = (X' X)^{-1} X' y \]

with \( E(b_t) = \beta \)

\[ V(b_t) = \sigma^2 (X' X)^{-1}. \]

To compare \( b_F \) and \( b_t \), we use the following result.

**Result:** If \( A \) and \( B \) are two positive definite matrices then \( A - B \) is at least positive semi definite if \( B^{-1} - A^{-1} \) is also at least positive semi definite.

Let

\[
A = (X' \bar{H}_z X)^{-1} \\
B = (X' X)^{-1} \\
B^{-1} - A^{-1} = X' X - X' \bar{H}_z X \\
= X' X - X' X + X' Z (Z' Z)^{-1} Z' X \\
= X' Z (Z' Z)^{-1} Z' X
\]

which is at least positive semi definite matrix. This implies that the efficiency declines unless \( X' Z = 0 \). If \( X' Z = 0 \), i.e., \( X \) and \( Z \) are orthogonal, then both are equally efficient.

The residual sum of squares under false model is

\[ SS_{res} = e_F' e_F \]

where

\[
e_F = y - Xb_F - Zc_F \\
b_F = (X' \bar{H}_z X)^{-1} X' \bar{H}_z y \\
c_F = (Z' Z)^{-1} Z' y - (Z' Z)^{-1} Z' X b_F \\
= (Z' Z)^{-1} Z' (y - X b_F) \\
= (Z' Z)^{-1} Z' \left[ I - X (X' \bar{H}_z X)^{-1} X' \bar{H}_z \right] y \\
= (Z' Z)^{-1} Z' \bar{H}_{xz} y \\
\bar{H}_z = I - Z (Z' Z)^{-1} Z' \\
\bar{H}_{xz} = I - X (X' \bar{H}_z X)^{-1} X' \bar{H}_z \\
\bar{H}_{xz}^2 = \bar{H}_{xz} \text{ : idempotent.}
\]
So

\[ e_F = y - X(X' \tilde{H}_Z X)^{-1} X' \tilde{H}_Z y - Z(Z' Z)^{-1} Z' \tilde{H}_Z y \]
\[ = \left[ I - X(X' \tilde{H}_Z X)^{-1} X' \tilde{H}_Z - Z(Z' Z)^{-1} Z' H_Z \right] y \]
\[ = \tilde{H}_Z y \]
\[ = \tilde{H}_Z^* y \text{ where } \tilde{H}_Z^* \equiv \tilde{H}_Z \tilde{H}_Z. \]

Thus

\[ SS_{res} = e_F e_F^\prime \]
\[ = y' \tilde{H}_Z \tilde{H}_Z^* \tilde{H}_Z^* \tilde{H}_Z y \]
\[ = y' \tilde{H}_Z \tilde{H}_Z^* \tilde{H}_Z^* y \]
\[ = y' \tilde{H}_Z^* y \]
\[ E(SS_{res}) = \sigma^2 \text{ tr}(\tilde{H}_Z^*) \]
\[ = \sigma^2 (n - k - r) \]
\[ E\left( \frac{SS_{res}}{n - k - r} \right) = \sigma^2. \]

So \( \frac{SS_{res}}{n - k - r} \) is an unbiased estimator of \( \sigma^2 \).

A comparison of exclusion and inclusion of variables is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Exclusion type</th>
<th>Inclusion type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation of coefficients</td>
<td>Biased</td>
<td>Unbiased</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Generally declines</td>
<td>Declines</td>
</tr>
<tr>
<td>Estimation of disturbance term</td>
<td>Over-estimate</td>
<td>Unbiased</td>
</tr>
<tr>
<td>Conventional test of hypothesis and confidence region</td>
<td>Invalid and faulty inferences</td>
<td>Valid though erroneous</td>
</tr>
</tbody>
</table>